文章编号　1004-924X(2026)16-2559-19

面向计算机辅助诊断的四叉树 token 分裂-合并去偏方法

周　涛[1,2], 沈飞宇[1,2*], 刘　洋[1,2], 常德芳[1,2], 于诗怡[1,2], 陈　铮[1,2]
(1. 北方民族大学 计算机科学与工程学院,宁夏 银川 750021;
2. 北方民族大学 图像图形智能处理国家民委重点实验室,宁夏 银川 750021)

摘要:在医学图像计算机辅助诊断(CAD)中,深度学习模型利用病灶区域与背景区域的虚假相关性进行预测,从而导致模型泛化能力与鲁棒性下降。为了解决这个问题,本文提出一种面向计算机辅助诊断的四叉树 token 分裂-合并去偏方法 QSMFormer。其主要贡献如下:首先,设计空间与语义耦合的异质性评估模块。通过联合建模区域显著性分布差异与语义一致性,自适应评估区域复杂性。其次,构建基于区域异质性的四叉树递归 token 分裂方法,对结构复杂区域进行细粒度划分,以保留关键病灶特征。最后,搭建语义与显著性感知的区域 token 合并模块,通过 token 间相似性与显著性动态合并相邻区域,采用 Top-K 策略筛选关键 token,从而减少冗余背景信息并增强区域语义表达的一致性。该方法在 Kaggle 胸部 X 光数据集与 Sharmin 9+1 数据集(一个十分类眼科数据集)上进行了验证。其中在 Kaggle 胸部 X 光数据集上取得 98.51% 的准确率和 97.02% 的 F1 分数;在 Sharmin 9+1 数据集上的准确率和 F1 分数分别达到 98.26% 和 92.60%。实验结果表明,QSMFormer 能够在提升分类性能的同时增强模型对背景扰动的稳定性,并在一定程度上缓解由非主体背景信息引起的潜在虚假关联影响。

关　键　词:医学图像分类;潜在伪相关缓解;Vision Transformer;四叉树 token 分裂;区域 token 合并
中图分类号:TP391.4　　**文献标识码**:A
doi:10.37188/OPE.20263416.2559　　**CSTR**:32169.14.OPE.20263416.2559

Quadtree token splitting-merging debiasing method for computer-aided diagnosis

ZHOU Tao[1,2], SHEN Feiyu[1,2*], LIU Yang[1,2], CHANG Defang[1,2], YU Shiyi[1,2], CHEN Zheng[1,2]

(1. *School of Computer Science and Engineering, North Minzu University, Yinchuan* 750021, *China*;
2. *Key Laboratory of Image and Graphics Intelligent Processing of State Ethnic Affairs Commission, North Minzu University, Yinchuan* 750021, *China*)
* *Corresponding author, E-mail: shenfeiyu*16*@stu. nmu. edu. cn*

Abstract: In medical image computer-aided diagnosis (CAD), deep learning models make predictions based on the spurious correlation between lesion areas and background areas, which leads to a decline in the model's generalization ability and robustness. To address this issue, the paper proposed a quad-tree token splitting-merging debiasing method called QSMFormer, which was specifically designed for computer-

收稿日期:2026-05-25;**修订日期**:2026-06-25.
基金项目:国家自然科学基金(No. 62576009, No. 62561002);宁夏自然科学基金(No. 2025AAC020006);宁夏领军人才培养项目(No. 2025GKLRLX26)

aided diagnosis. Its main contributions were as follows: Firstly, designing a spatial-semantic coupled heterogeneity evaluation module. By jointly modeling the differences in regional significance distribution and semantic consistency, the complexity of the region could be adaptively evaluated. a quadtree-based recursive token splitting method driven by region heterogeneity is constructed, which performed fine-grained partitioning on structurally complex regions to preserve critical lesion features. Finally, building a semantic- and saliency-aware region token merging module. By dynamically merging adjacent regions based on the similarity and significance between tokens, and using the Top-K strategy to select key Tokens, redundant background information was reduced and the consistency of regional semantic expression was enhanced. This method was verified on the Kaggle Chest X-ray dataset and the Sharmin 9+1 dataset (a ten-class ophthalmology dataset). Specifically, on the Kaggle Chest X-ray dataset, an accuracy rate of 98.51% and an F1 score of 97.02% were achieved; on the Sharmin 9+1 dataset, the accuracy rate and F1 score reached 98.26% and 92.60% respectively. The experimental results show that QSMFormer improves classification performance and enhances stability under background perturbations, while reducing the influence of non-subject background information to some extent.

Key words: medical image classification; potential spurious correlation mitigation; vision transformer; quadtree token splitting; region token merging

1 引　言

近年来，计算机辅助诊断(Computer-Aided Diagnosis, CAD)对模型在细粒度病灶识别、结构化特征提取以及跨域稳健性方面有更高要求[1]。视觉Transformer(ViT)凭借其全局自注意力机制和强大的特征建模能力，逐渐成为满足这些需求的重要架构[2]。在CAD场景中，ViT采用基于token的表征方式，使其在捕获病灶的细粒度结构与长程依赖关系方面展现出显著的结构性优势[3]。然而，ViT在高分辨率医学影像上通常需要处理大量token，不仅带来巨大的计算与内存开销，更重要的是，这些token中往往包含大量与诊断无关的背景信息[4]。模型在训练过程中可能会错误利用这些不相关背景特征，从而建立诊断结果和背景特征的虚假关联(spurious correlations)[5]，导致诊断偏差，并降低模型在背景变化或外部数据上的稳定性。因此，设计一种高效且能够降低背景冗余信息干扰的token选择机制，对CAD任务尤为关键[6]。

为降低计算开销并缓解虚假关联问题，已有研究提出了多种token剪枝与token合并策略[7]。Rao等[8]提出DynamicViT，通过一个轻量级的MLP模块预测每个token的保留概率，并利用Gumbel-Sigmoid技巧对概率进行二值采样，生成二进制掩码以筛选出语义信息更丰富的token，同时保持端到端的可训练性。Bolya等[9]提出Token Merging(ToMe)方法，使用通用且无需训练的轻量级匹配算法，在Transformer的每一层中逐步识别并合并相似token，从而在保持较高精度的前提下显著减少token数量，加速模型推理。Liang等[10]依据Class token的注意力权重，通过合并非关键token来重组图像序列，从而实现加速。Zeng等[11]采用基于K近邻密度峰值的聚类算法对token进行分簇，并对同一簇内的token进行加权平均合并。李敏等[12]提出一种基于相互注意力权重选择(Mutual Attention Weight Selection, MAWS)的token选择方法，通过对多层Encoder收集的特征token进行筛选，选出最有助于胸部X光多标签分类的特征。然而，这些方法在token选择过程中仍易引入由背景或无关特征驱动的选择偏差。

本文受图像分割领域中区域分裂与合并思想的启发，设计了一种基于图像块的四叉树区域分裂-合并方法。具体而言，首先，提出空间与语义耦合的异质性评估模块，通过联合建模区域内部的显著性离散程度与语义一致性，自适应评估不同区域的结构复杂性，为后续区域划分提供依据。其次，提出基于区域异质性的四叉树递归token分裂方法，对内部结构复杂或语义变化显著

的区域进行递归细粒度划分，从而获得不同尺度的 patch 区域，并充分保留具有诊断价值的局部信息。最后，提出语义与显著性感知的区域 token 合并模块，根据区域间的嵌入相似性与 token 显著性一致性，对相邻且语义相关的区域进行动态合并，并从各区域中筛选最具判别能力的关键 token，在减少冗余背景干扰的同时重建稳定的区域级语义结构，从而提升模型对关键目标区域的感知能力与判别性能。

本文的主要贡献可以概括为以下四点：

（1）提出面向计算机辅助诊断的四叉树 token 分裂-合并去偏方法 QSMFormer。该方法从区域结构出发，采用“异质性评估—分裂—合并”机制，聚焦于不规则区域中的关键特征，从而减少固定尺度划分带来的背景冗余信息干扰。

（2）设计空间与语义耦合的异质性评估模块，对各区域内部的特征响应分布与语义结构进行联合建模与量化。该方法通过分析区域内 token 的显著性差异与语义相似性分布，判别高异质（结构复杂、语义不一致）与低异质（语义稳定）区域，从而为后续区域分裂与合并提供明确的决策依据。该机制有助于模型区分具有诊断意义的关键区域与背景干扰区域，从特征选择层面降低对非主体背景信息的依赖。

（3）设计基于区域异质性的四叉树递归 token 分裂方法，该方法通过四叉树分裂机制，对内部结构复杂或语义变化显著的区域进行四叉树细粒度划分，从而保留关键的局部诊断信息。相比于固定粒度建模，该策略能够动态调整区域尺度，实现更加精细的空间表征，从而降低跨区域噪声被误作为稳定判别特征的可能性。

（4）提出语义与显著性感知的区域 token 合并模块，该方法将语义与显著性一致的相邻区域归并为若干不规则区域，再使用 top-k 策略从区域中选择最有价值的 token，在减少冗余 token 的同时保留空间多样性与关键诊断信息，进而在一定程度上缓解潜在伪相关信息对分类决策的影响。

2　研究动机与问题分析

Vision Transformer 中的 token pruning 是指在 ViT 模型推理过程中，动态移除图像中冗余的图像块（patch）对应的 token，从而减少计算量、加速推理的技术。Token Pruning 通过直接丢弃 token 实现信息删除，导致保留的 token 在空间上形成不连续的稀疏网格，破坏了图像的局部结构连续性。Token Pruning 的剪枝过程如图 1（a）所示（彩图见期刊电子版），在图 1（a）中，与雪兔相关的 token 保留了 8 个（用绿色√表示），5 个包含有雪兔的耳朵和腿的图像块被误分到背景中（用黄色×表示），同时，存在 5 个背景区域被误分到雪兔中（用红色×表示）。按图中标注统计，Token pruning 对目标相关 token 的保留比例为 44.44%。该示意结果表明，直接剪枝可能造成目标区域 token 的遗漏以及背景 token 的误保留，从而增加模型利用非目标区域信息进行判断的风险。

Token Merging 则是一种自下而上的聚合策略，通过将局部语义信息相近的 token 合并为更大的语义区域以压缩冗余信息。Token Merging 通过保持空间连续性的局部合并机制来减少 token，但大背景中的小目标易被相邻或相似的背景 token 合并吸收，导致细节信息丢失。Token Merging 的剪枝过程如图 1（b）所示，在图 1（b）中，与雪兔相关的 token 保留了 8 个（用绿色√表示），4 个包含雪兔的耳朵和腿的图像块被误认定为背景区域并且与背景合并（用红色×表示），同时，存在 1 个背景区域块被认定为雪兔并且与雪兔合并（用黄色×表示）。按图中标注统计，Token merging 对目标相关 token 的保留比例为 61.54%。该示意结果表明，基于相似性的合并策略可能将目标边缘区域与相邻背景区域合并，从而导致细粒度诊断信息被削弱。

本文受图像分割的像素级区域分裂合并方法的启发，设计基于图像块的四叉树区域分裂合并方法，具体思路如图 1（c）所示。在图 1（c）中，包括三个阶段。第一个阶段采用图像块级的四叉树分裂策略，通过异质性评估策略进行嵌套分解，得到了不同大小的 patch 块。第二个阶段采用了区域合并策略，将相似的不同尺寸的 patch 块进行合并，把图像分为数个大小不一的不规则区域。第三个阶段采用区域选择策略，选择每个区域最重要的 token。按图中标注统计，该示意样例中目标相关 patch 均被保留，说明分裂-合并策略有助于维持目标区域的空间完整性。

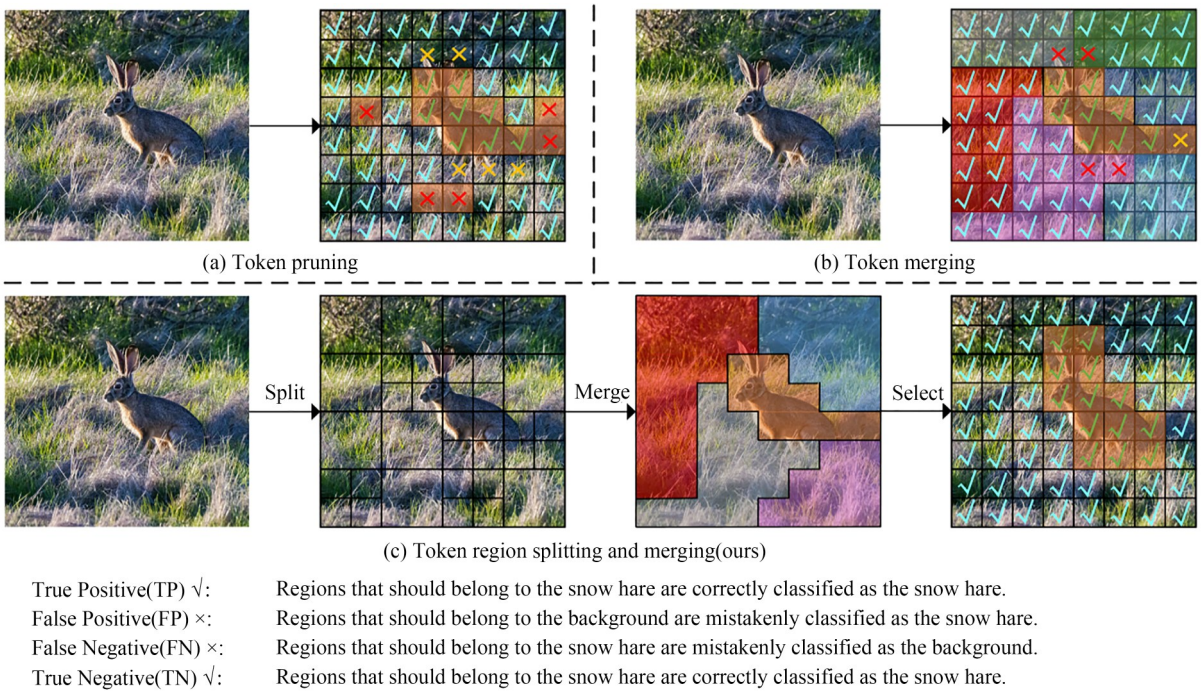

图 1　不同 token 选择策略在空间结构保持与语义选择上的对比

Fig. 1　Comparison of different token selection strategies in terms of maintaining spatial structure and semantic selection

3　相关工作进展

3.1　医学图像识别中的去偏见问题

医学图像识别中的偏见问题，指的是模型对不同偏见属性（如种族、性别、年龄等）的医学图像存在区别对待，进而在识别目标属性时，对不同群体的图像表现出显著差异的准确率。由于模型以最小化整体误差为优化目标，它会优先拟合多数群体图像以降低总误差，从而忽略少数群体。这种拟合策略导致模型在不同群体间的性能失衡，最终可能使某些群体或在特定条件下的患者面临误诊或被忽视的风险。这类问题可能会对患者健康产生不良影响，并加剧社会健康的不平等，因此去除“偏见”，即非歧视性的算法决策程序越来越重要[13]。Tejasvi[14]进一步讨论了机器学习公平性和可解释性在特定阶段癌症生存能力预测中的重要作用。Rubén[15]提出了一种新的衡量公平的指标，即平等和公平。

Vision Transformer（ViT）[2]是一种基于 Transformer 架构的图像分类模型。其核心思想是将输入图像划分为一系列固定大小的图像块（patches），将每个块展平后送入 Transformer 模型进行处理。与依赖卷积核局部感受野的传统方法不同，ViT 的创新之处在于利用自注意力机制（self-attention）直接捕捉图像中的全局信息，从而能够在更大尺度上学习图像的长距离依赖关系。Evgin Goceri[16]提出了一个具有 CNN 和 Transformer 并行结构的网络 FairDRO，将重新加权和正则化结合起来，采用一个公平指标去偏见。Seyyed-Kalantari L[17]讨论了因为样本不足导致的胸片诊断偏见问题。因此，如何减少医学图像模型对数据偏倚和非诊断相关特征的依赖，是提升模型可靠性的重要问题。

3.2　Token 选择策略：剪枝与合并

在 Vision Transformer 的相关研究中，token 的选择策略对模型的效率与性能至关重要。当前，降低 token 冗余的主流方法主要分为剪枝（pruning）与合并（merging）两类。

Token 剪枝通过丢弃冗余或信息量低的 token，保留最具判别力的部分。其优点在于能显著减少计算量并加速推理，缺点则是可能丢失局部细节或边缘信息，尤其在深层网络中易导致不可逆的信息损失。例如，Dynamic ViT[8]在推理过程中逐步丢弃不重要的 token，而 EViT[10]则利用注意力得分来识别并剔除冗余 token。

与之相对，token 合并策略并非直接丢弃 token，而是将相似的 token 融合为一个新 token，从而在降低序列长度的同时保留更多原始信息。

该方法能更好地维持特征完整性，但实现复杂度更高，且可能因过度融合而模糊不同语义区域的边界。典型工作包括 Token Merging（ToMe）[9]，它通过二分图软匹配在合并过程中有效平衡效率与精度，以及 TokenPool[18]，它利用可学习的聚类机制将相似 token 动态聚合。总体而言，剪枝更关注效率提升，而合并在信息保留方面具有天然优势。

3.3　图像分割中的分裂与合并方法

在图像分割领域，分裂与合并（Split-and-Merge）方法[19]是一种经典的区域划分策略，通过结合区域分裂与区域归并两种机制，实现对图像结构的建模。该方法通常建立在金字塔或四叉树等层级表示上，首先定义初始区域划分及一致性判定准则。随后在分裂阶段，自顶向下递归处理：当某一区域内部特征（如灰度或纹理）不满足一致性条件时，将其划分为多个子区域，从而逐步细化图像表示。

在完成局部分裂后，方法会对相邻区域进行归并判断：若子区域之间或相邻区域之间满足一致性约束，则将其合并为更大区域，以减少过度分割带来的碎片化问题。同时，一些方法还支持跨层级的归并操作，使得不同尺度下的区域能够进一步整合。最后，通过对过小区域进行邻域合并优化，提升分割结果的稳定性与语义完整性。总体而言，该方法通过“先分裂、后归并”的策略，在细节保留与全局一致性之间取得了良好平衡。

4　原理与算法

针对计算机辅助诊断中背景冗余信息易干扰模型判别的问题，本文提出一种基于四叉树区域分裂与合并的 token 选择方法，称为 QSMFormer。

QSMFormer 的整体结构如图 2 所示。本文以 12 层 ViT-Base 作为骨干网络，在中间层（第 7

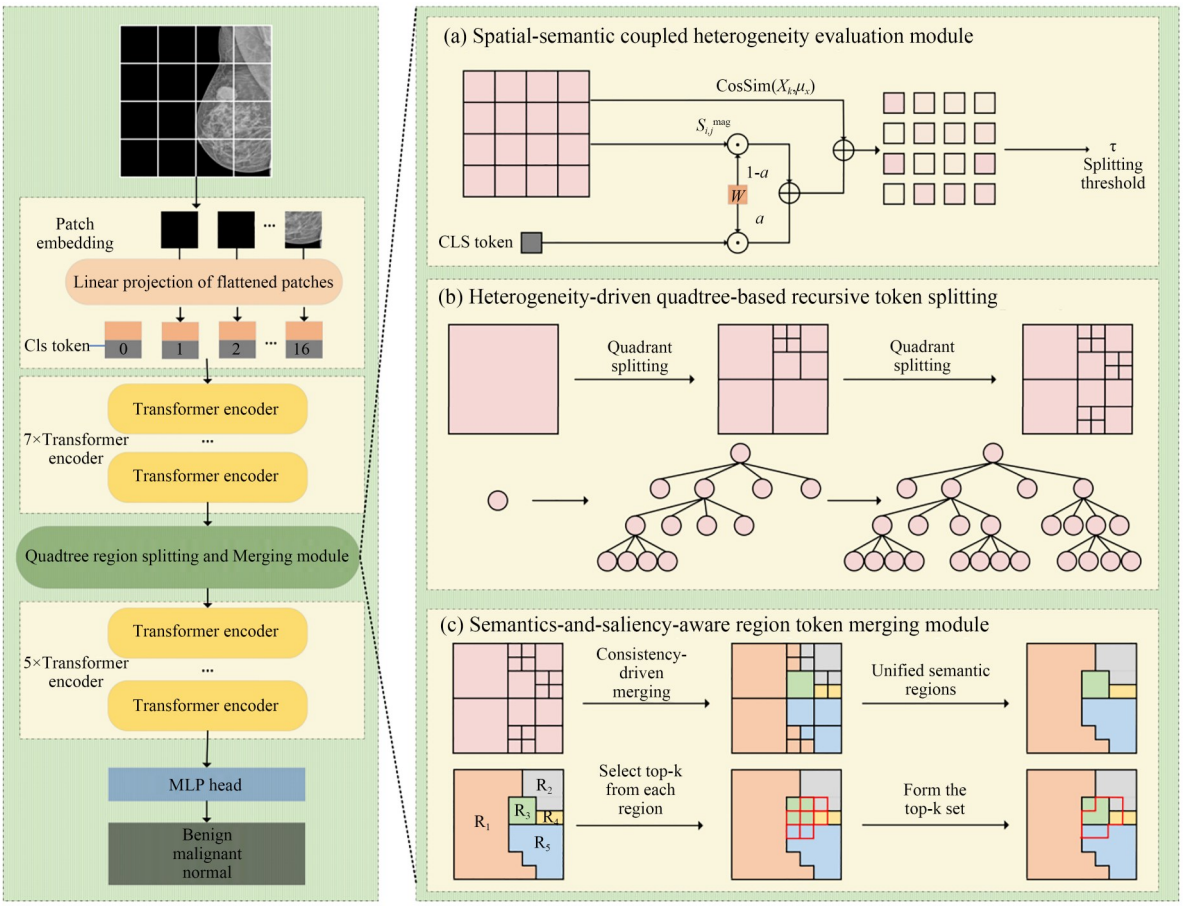

图 2　QSMFormer 总体结构图

Fig. 2　Overall architecture of QSMFormer

个 Transformer Encoder 之后）插入所提出的 QSM 模块，以实现 token 选择。具体而言，输入图像首先经过前 7 层 Transformer Encoder 提取中间特征表示，随后进入 QSM 模块进行结构化优化，最后通过剩余 5 层 Transformer Encoder 进一步建模高层语义并完成分类任务。

QSM 模块由三个关键子模块组成：(A)空间与语义耦合的异质性评估模块、(B)基于区域异质性的四叉树递归 token 分裂模块以及(C)语义与显著性感知的区域 token 合并模块。该模块通过对 token 特征进行空间重组与区域级建模，实现从“逐 token 独立选择”向“区域感知选择”的转变，从而在减少冗余 token 的同时保留关键诊断信息。

4.1　空间与语义耦合的异质性评估模块

为降低模型对非主体背景区域的过度依赖，本节设计空间与语义耦合的异质性评估模块，对输入的 token 特征表示的空间结构复杂度进行建模。如图 3 所示，该模块通过联合建模 token 显著性与区域内部语义一致性，构建统一的区域异质性评价策略，并生成区域分裂的全局阈值。

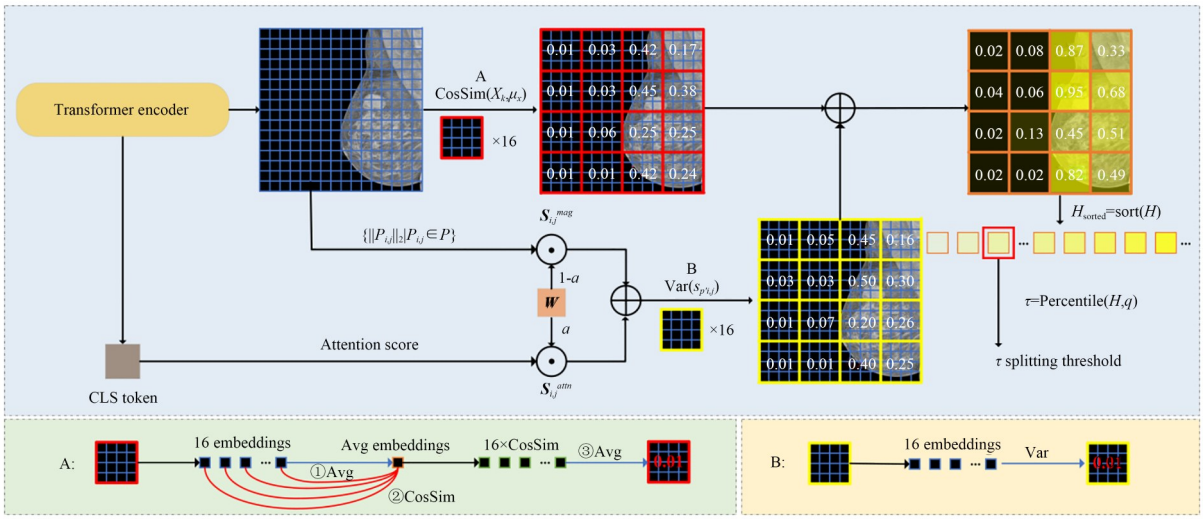

图 3　空间与语义耦合的异质性评估模块

Fig. 3　Spatial-semantic coupled heterogeneity evaluation module

具体而言，首先将输入的 token 特征按照其对应的空间位置组织为 16×16 的 patch 网格，形成一个规则的 patch 集合：

$$P=\left\{p_{i,j} | i=1,2,\cdots,I_p, j=1,2,\cdots,J_p\right\}, \quad (1)$$

其中：I_p 和 J_p 分别表示 patch 网格在行和列方向上的划分数量。每个 patch 对应一个 token 表征。

对于每个 patch$p_{i,j}$，利用其对应的 token embedding $x_{i,j}$ 的 L2 范数来衡量局部特征强度，其定义为：

$$s_{i,j}^{\mathrm{mag}}=\left\| x_{i,j}\right\|_2=\sqrt{x_1^2+x_2^2+\cdots+x_D^2}, \quad (2)$$

其中，$x_{i,j}\in\boldsymbol{R}^D$ 表示该 patch 的特征向量，每一维度对应不同的语义或纹理通道响应。

与此同时，从 Vision Transformer 的 CLS Token 中提取其对各 patch token 的注意力权重。自注意力机制中，query 与 key 之间的相似性通过点积计算：

$$Attn(\boldsymbol{Q},\boldsymbol{K})=\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}, \quad (3)$$

其中：$\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 分别表示 query 和 key 向量，d_k 为 key 的维度。随后，对注意力得分进行 softmax 归一化，得到每个 token 的注意力权重：

$$s_{i,j}^{\mathrm{attn}}=softmax\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}}\right). \quad (4)$$

最终，通过线性加权的方式融合幅值分数与注意力分数，得到 token 级显著性分数：

$$s_{i,j}=\alpha s_{i,j}^{\mathrm{attn}}+(1-\alpha)s_{i,j}^{\mathrm{mag}}, \quad (5)$$

其中，$\alpha\in[0,1]$ 为融合权重超参数。由此得到的

显著性分数向量：

$$s=\{s_{i,j}|p_{i.j}\in P\}\in \boldsymbol{R}^{N}. \tag{6}$$

在获得每个 token 的显著性分数后，本文进一步利用粗粒度网格划分对图像的区域异质性进行建模。具体而言，将输入图像按照宽度 $W/2$ 与高度 $H/2$ 划分为 4×4 个子区域集合：

$$P'=\{p'i,j|i,j\in\{1,2,3,4\}\}, \tag{7}$$

其中，每个 $p'_{i,j}$ 表示一个空间上不重叠的子区域。

对于每个子区域 $p'_{i,j}\in P'$，首先计算区域内所有 token 的 embedding 与区域均值向量之间的余弦距离，以评估区域内部的语义一致性。余弦相似度的定义如下：

$$CosSim(u,v)=\frac{u\cdot v}{\|u\|\|v\|}, \tag{8}$$

其中：u 和 v 表示两个特征向量，$\|\cdot\|$ 为 L2 范数。基于此，余弦距离定义为：

$$CosDist(u,v)=1-CosSim(u,v). \tag{9}$$

区域 $p'_{i,j}$ 的均值向量 $\mu_x^{(p'_{i,j})}$ 代表该区域内所有 token embedding 的中心特征，可通过式(10)获得：

$$\mu_x^{(p'_{i,j})}=\frac{1}{|p'_{i,j}|}\sum_{k\in p'_{i,j}}^{N}X_k, \tag{10}$$

其中：X_k 表示区域 $p'_{i,j}$ 内第 k 个 token 的 embedding，$|p'_{i,j}|$ 为该区域内 token 的数量。

据此，区域 $p'_{i,j}$ 内所有 token 与均值向量之间的平均余弦距离定义为：

$$AvgCosDistToMean\left(X_{p'_{i,j}}\right)=\frac{1}{|p'_{i,j}|}\sum_{k\in p'_{i,j}}CosDist(x_k,\mu_x^{(p'_{i,j})}). \tag{11}$$

该度量用于反映区域内部 token 表征的一致性程度。平均余弦距离越大，表明区域内 token 差异性越强、语义结构越复杂。

与此同时，计算区域内 token 显著性分数的方差以刻画显著性离散程度：

$$Var\left(s_{p'_{i,j}}\right)=\frac{1}{|p'_{i,j}|}\sum_{k\in p'_{i,j}}(s_k-\mu_s^{(p'_{i,j})})^2, \tag{12}$$

其中：s_k 表示第 k 个 token 的显著性分数，$\mu_s^{(p'_{i,j})}$ 为该区域 token 分数的均值。

对于每一个子区域 $p'_{i,j}\in P'$，综合区域内 token 显著性分数的离散程度与语义不一致性，定义其异质性分数为：

$$H_{p'_{i,j}}=Var\left(s_{p'_{i,j}}\right)+\lambda\cdot AvgCosDistToMean(X_{p'_{i,j}}), \tag{13}$$

其中：$\mathrm{Var}\left(s_{p'_{i,j}}\right)$ 表示子区域 $p'_{i,j}$ 内所有 token 显著性分数的方差，$\mathrm{AvgCosDistToMean}(X_{p'_{i,j}})$ 表示该区域内 token embedding 与区域均值向量之间的平均余弦距离。λ 为平衡参数，用于控制显著性离散性与语义不一致性在整体异质性中的相对权重。

最后，对所有子区域的异质性分数进行升序排序，并选取第 q 百分位对应的数值作为全局分裂阈值：

$$\tau=Percentile(H,q), \tag{14}$$

本文默认 $q=65$ 该阈值能够反映图像整体复杂程度，并作为后续四叉树区域分裂过程的判定依据。

综上，所提出的模块通过统一建模 token 级显著性与区域级语义一致性，从多个空间尺度刻画图像结构复杂性，为后续的区域分裂与合并提供了稳定且有效的全局指导信号。

4.2　基于区域异质性的四叉树递归 token 分裂模块

在获得全局异质性阈值 τ 后，本文进一步设计基于区域异质性的四叉树递归 token 分裂模块，以根据区域内部结构复杂度对图像进行层级化空间划分，从而降低跨区域背景信息与主体结构信息被混合建模的风险。如图 4 所示，该模块在全局阈值约束下，通过递归式分裂构建与图像内容复杂度相匹配的非均匀区域结构。

具体而言，区域分裂以整幅图像作为根节点，采用四叉树结构进行递归划分。对于任意候选区域 R，首先基于第 4.1 节定义的异质性度量函数计算其异质性分数：

$$H(R)=Var(S_R)+\lambda\cdot AvgCos(X_R,\mu_R), \tag{15}$$

该指标综合刻画了区域内显著性分布的离散程度与语义表示的不一致性。

基于该异质性度量，区域是否继续分裂由如下准则决定：当 $H(R)>\tau$ 时，表明该区域内部存在显著的结构变化或语义不一致性，需要进一步细化表示；否则认为该区域相对均质，可保持当前尺度并作为叶节点保留。

当满足分裂条件时，区域 R 将沿空间维度被

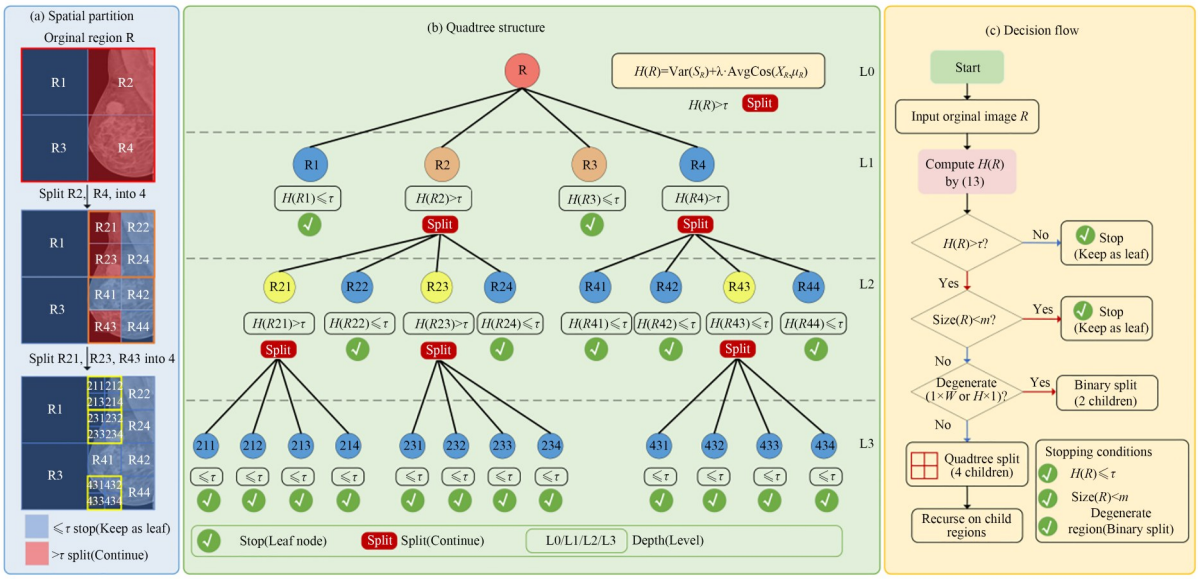

图 4　基于区域异质性的四叉树递归 token 分裂模块

Fig. 4　Heterogeneity-driven quadtree-based recursive token splitting module

划分为四个子区域{R1,R2,R3,R4},并对每个子区域递归执行相同的异质性评估与分裂过程,从而形成层级化的空间划分结构。该过程可形式化为如下递归函数:

$$L = Split(R(0), 0), \tag{16}$$

其中:$R(0)$表示初始区域,L为最终得到的叶节点区域集合。

为了保证划分过程的稳定性与可控性,区域分裂受到以下停止条件约束:(1)当区域异质性低于阈值τ时,停止分裂;(2)当区域尺寸小于预设最小尺度m时,停止分裂;(3)当区域在空间上退化为单行或单列结构时,采用二叉划分以避免不合理的四叉分裂。上述机制确保了划分过程既能够充分细化复杂区域,又避免在简单区域中产生不必要的计算开销。

从结构上看,如图4中部所示,整个分裂过程可以表示为一棵深度不均的四叉树,不同分支根据区域异质性展开,形成非均匀的层级结构。图4右侧进一步给出了对应的决策流程,清晰地展示了从异质性计算、阈值判断到递归分裂的完整过程。

通过该异质性驱动的递归划分机制,模型能够在整幅图像上构建与内容复杂度相匹配的空间结构:语义变化剧烈或纹理复杂的区域被细化为高分辨率子区域,而结构简单或语义一致的区域保持较大尺度。这种非均匀划分方式不仅避免了均匀网格划分带来的冗余计算,还为后续区域合并与多尺度特征建模提供了结构化基础。区域分裂的具体实现细节如算法1所示。

Algorithm 1　Region splitting based on heterogeneity

Input: Token scores $s=[s_1,\cdots,s_N]^{\mathrm{T}}\in\boldsymbol{R}^N$, token embeddings $\boldsymbol{X}=[X_1,\cdots,X_N]^{\mathrm{T}}\in\boldsymbol{R}^{N\times D}$, splitting threshold $\tau>0$, embedding weight $\lambda\geqslant 0$, minimum region size $m\in\boldsymbol{N}^+$, numerical constant $\epsilon>0$

Output: Set of leaf regions L

Initialize root region $R^{(0)}$ covering all N tokens

($L\leftarrow$ Split($R^{(0)}$, 0)

return L

function Split(R, d)

if$|R|\leqslant m$ then //Region too small to split

return$\{R\}$

end if

Compute region mean embedding:

$$\mu\leftarrow\frac{1}{|R|}\sum_{i\in R}X_i$$

Compute region mean score:

$$\bar{s}\leftarrow\frac{1}{|R|}\sum_{i\in R}S_i$$

Compute heterogeneity score:

$$\mathrm{H}(R)\leftarrow\frac{1}{|R|}\sum_{i\in R}\|x_i-\mu\|^2+\lambda\cdot\frac{1}{|R|}\sum_{i\in R}(s_i-\bar{s})^2$$

14：if $H(R)\leqslant\tau+\epsilon$ then //Homogeneous region
15：return $\{R\}$
16：end if
17：Split R into four child regions
18：$\{R_1, R_2, R_3, R_4\}$ using quadtree partitioning
19：$L'\leftarrow\varnothing$
20：for k=1 to 4 do
21：$L'\leftarrow L'\cup\text{Split}(R_k, d+1)$
22：end for
23：return L'

24：end function

4.3　语义与显著性感知的区域 token 合并模块

为缓解递归分裂可能带来的区域碎片化问题，并避免语义不一致区域被错误聚合，本节提出语义与显著性感知的区域 token 合并模块。该名称中的“语义感知”指通过区域特征向量表征刻画高层语义一致性，“显著性感知”则通过 token 分数反映区域的重要性水平。二者的联合建模使得区域合并不仅关注特征相似性，还兼顾信息贡献度，从而避免将语义不同或重要性差异较大的区域错误合并。

该模块的核心目标是：在保证空间连续性的前提下，将分裂阶段得到的碎片化叶节点区域自适应整合为语义一致且结构紧凑的区域集合。如图 5 所示，区域合并仅在空间相邻区域之间进行，并通过语义一致性与显著性一致性的双重约束判定区域间的可合并性。相比仅基于特征相似性的合并方法，该模块具有以下特点：(1)双重一致性约束：同时考虑 embedding 相似性与显著性差异，提升合并可靠性；(2)空间约束性：仅在相邻区域间执行操作，保证结构合理性；(3)全局一致性聚合：基于并查集实现跨区域传递式合并，形成稳定区域簇。具体过程包括以下两个步骤：

步骤一对区域进行一致性度量。给定分裂阶段得到的叶节点区域集合：

$$L=\{R_i\}. \tag{17}$$

首先计算每个区域的语义表示与显著性统计量。对于区域 R_i，其均值 embedding 与平均显著性分数分别定义为：

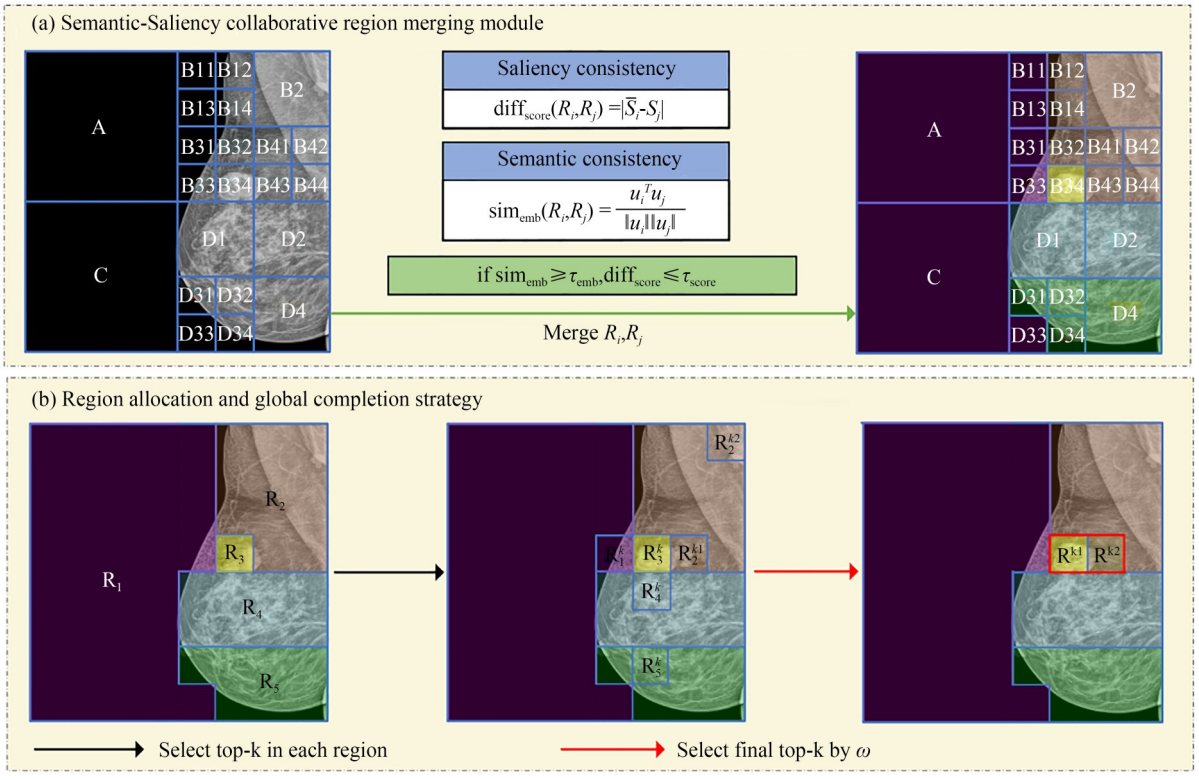

图 5　语义与显著性感知的区域 token 合并模块

Fig. 5　Semantics-and-saliency-aware region token merging module

$$u_i=\frac{1}{|R_i|}\sum_{x_k\in R_i}x_k, \tag{18}$$

$$\bar{s}_i=\frac{1}{|R_i|}\sum_{x_k\in R_i}s_k. \tag{19}$$

对于任意空间相邻区域对(R_i,R_j)，计算其语义相似性与显著性差异：

$$sim_{\text{emb}}(R_i,R_j)=\frac{u_i^T u_j}{\|u_i\|\|u_j\|}, \tag{20}$$

$$diff_{\text{score}}(R_i,R_j)=\left|\bar{s}_i-\bar{S}_j\right|. \tag{21}$$

该步骤从语义空间与重要性分布两个角度刻画区域间的一致性，为后续合并决策提供判据。

步骤二基于一致性对区域进行聚合。在获得区域间一致性度量后，依据如下准则执行区域合并：

$$sim_{\text{emb}}\geqslant\tau_{\text{emb}},\ diff_{\text{score}}\leqslant\tau_{\text{score}}. \tag{22}$$

当满足上述条件时，认为两个区域在语义表达与显著性水平上具有一致性，将其合并为同一结构单元。为高效实现多区域聚合，本文采用并查集(Union-Find)结构对满足条件的相邻区域进行动态合并，从而形成若干区域簇。

在完成所有合并操作后，每个区域簇通过其最小外接矩形进行统一表示，得到新的区域集合M。该过程能够整合分裂阶段产生的冗余子区域，使得空间划分结果在语义上更加连贯、在结构上更加紧凑，同时为后续的区域级建模与token选择提供更加稳定的空间基础。

在完成区域整合后，进一步在预算约束K下对token进行筛选，以在控制计算开销的同时最大化信息表达能力。该过程在区域结构约束与全局最优目标之间进行协同优化，统一实现区域均衡性与全局一致性。

具体而言，首先根据合并后区域的重要性与空间覆盖范围，为每个区域分配token配额。对于区域R_i，其权重由区域平均显著性分数与区域面积共同决定：

$$\omega=(mean\_score_i)^p\times(area_i)^q, \tag{23}$$

其中：mean_score_i表示区域内token显著性分数的均值，area_i表示区域空间范围，p与q为调节参数。

基于权重比例，采用最大剩余法将总预算K分配为整数配额：

$$k_i\propto\frac{\omega_i}{\sum_j\omega_j},\ \sum_i k_i=K. \tag{24}$$

在获得各区域配额后，于每个区域内部执行Top-k_i筛选：若区域内token数量不超过配额，则全部保留；否则选取得分最高的k_i个token，从而优先保留区域内最具代表性的局部信息。

然而，区域级筛选后得到的token总数在实际中可能与预算K存在偏差。为此，引入全局一致性校准机制对结果进行统一调整：当筛选后的token数量少于K时，从未被选中的token集合中按照显著性分数由高到低依次补充，直至达到预算；反之，当token数量超过K时，对当前候选集合执行全局Top-K裁剪，仅保留得分最高的token。

通过上述“区域配额约束＋全局一致性校准”的协同机制，最终得到满足预算约束的token集合。该过程在保证空间分布均衡性的同时，实现了对高价值信息的优先保留，从而获得更加紧凑且具有判别性的token表示。区域合并的整体流程如算法2所示。

Algorithm 2: Merging adjacent regions based on embedding and score similarity

Input: Set of leaf regions $\mathrm{L}=\{R_i\}_{i=1}^{M}$ from Algorithm 1, token embeddings $\boldsymbol{X}=[X_1,\cdots,X_N]^{\mathrm{T}}\in\boldsymbol{R}^{N\times D}$, token scores $s=[s_1,\cdots,s_N]^{\mathrm{T}}\in\boldsymbol{R}^N$, embedding similarity threshold $\tau_{\text{emb}}\in[0,1]$, score difference threshold $\tau_{\text{score}}>0$, numerical stability constant $\in>0$

Output: Merged regions $\mathrm{M}=\{(\text{top}_k,\text{left}_k,h_k,\omega_k)\}_{k=1}^{K}$, where each tuple represents a bounding box

1. Initialize Union-Find structure μ for all regions in L
2. for each region $R_i\in L$ do
3. Compute region mean embedding:
4. $\mu_i\leftarrow\frac{1}{R_i}\sum_{j\in R_i}X_j$
5. Compute region mean score:
6. $\bar{s}_i\leftarrow\frac{1}{R_i}\sum_{j\in R_i}s_j$
7. end for
8. for each pair of regions (R_i,R_j) where $i<j$ do
9. if AreaAdjacent(R_i,R_j) then
10. Compute embedding similarity:

11. $\mathrm{sim}_{\mathrm{emb}} \leftarrow \dfrac{\mu_i^{\mathrm{T}} u_j}{\left\| \mu_i \right\| \cdot \left\| \mu_j \right\|}$ //Cosine similarity
12. Compute score difference:
13. $\mathrm{diff}_{\mathrm{score}} \leftarrow \left| \bar{s}_i - \bar{s}_j \right|$
14. if $\mathrm{sim}_{\mathrm{emb}} \geqslant \tau_{\mathrm{emb}}$ and $\mathrm{diff}_{\mathrm{score}} \leqslant \tau_{\mathrm{score}} + \epsilon$ then
15. μ. Union(i, j) //Merge similar adjacent regions
16. end if
17. end if
18. end for
19. Group regions based on μ parent nodes $\rightarrow \{G_k\}_{k=1}^{K}$
20. $M \leftarrow \varnothing$
21. For each group G_k do
22. Compute bounding box coordinates:
23. $\mathrm{top}_k \leftarrow \min\{\mathrm{top}(R) | R \epsilon G_k\}$
24. $\mathrm{left}_k \leftarrow \min\{\mathrm{left}(R) | R \epsilon G_k\}$
25. $\mathrm{bottom}_k \leftarrow \max\{\mathrm{top}(R) + \mathrm{height}(R) | R \in G_k\}$
26. $\mathrm{right}_k \leftarrow \max\{\mathrm{left}(R) + \mathrm{width}(R) | R \in G_k\}$
27. Compute height and width:
28. $h_k \leftarrow \mathrm{bottom}_k - \mathrm{top}_k$
29. $w_k \leftarrow \mathrm{right}_k - \mathrm{left}_k$
30. $M \leftarrow M \cup \{(\mathrm{top}_k, \mathrm{left}_k, h_k, w_k)\}$
31. end for
32. return M
33. function AreAdjacent(R_i, R_j)
34. Let(t_i, l_i, h_i, w_i) and (t_j, l_j, h_j, w_j) be bounding boxes
35. return$(t_i \leqslant t_j + h_j$ and $t_j \leqslant t_i + h_i)$ and
36. $(l_i \leqslant l_j + w_j$ and $l_j \leqslant l_i + w_i)$ and
37. $(\max(t_i, t_j) < \min(t_i + h_i, t_j + h_j)$ or $\max(l_i, l_j) < \min(l_i + w_i))$
38. end function

5　实验结果与分析

5.1　数据集与预处理

数据集 1：卡塔尔大学、卡塔尔多哈和孟加拉国达卡大学的一组研究人员以及来自巴基斯坦和马来西亚的合作者与医生合作创建的胸部 X 光图像数据，该数据集在 Kaggle 上免费提供[20]。实验过程中共选取 5 374 张图像，其中包括 1 362 张正常图像，1 332 张新冠肺炎图像，1 335 张肺部空洞图像以及 1 345 张病毒性肺炎图像。将所有不同尺寸大小的原始图像缩放至 224×224 pixels，训练集和验证集的比例按照 9∶1 划分，然后将其转换为向量格式并对其进行像素值归一化处理。

数据集 2：该数据集由 Sharmin 等人创建，并在题为《用于眼科疾病检测和分类的彩色眼底图像数据集》的论文中进行了描述，[21]。数据在八个月内从 Anawara Hamida 眼科医院和 B. N. S. B. Zahurul Haque 眼科医院使用彩色眼底照相仪收集。它包括两种类型的图像：分为九类（糖尿病视网膜病变、青光眼、黄斑瘢痕、视盘水肿、中心性浆液性脉络膜视网膜病变（CSCR）、视网膜脱离、视网膜色素变性、近视和健康）的彩色眼底照片，以及分为一类（翼状胬肉）的前段图像。该数据集包含 5 335 张原始图像。本文将原始图像按 70%，20% 和 10% 的比例分为训练、验证和测试集。

5.2　评价指标以及复杂度计算

为了全面评估本文方法在分类任务中的性能，本文从分类性能和模型复杂度两个方面对模型进行评价。在分类性能方面，采用准确率（Accuracy，Acc）、精确率（Precision，Pre）、召回率（Recall，Rec），F1 Score 和特异性（Specificity，Spe）作为评价指标。根据图像真实类别与模型预测类别构建混淆矩阵，并通过混淆矩阵中的真正例（True Positive，TP）、假正例（False Positive，FP）、真负例（True Negative，TN）和假负例（False Negative，FN）计算上述指标，其计算公式如式（25）～式（29）所示：

$$Acc = \frac{TN + TP}{FP + TN + TP + FN}, \tag{25}$$

$$Pre = \frac{1}{c} \sum_{i=1}^{c} \frac{TP_i}{FP_i + TP_i}, \tag{26}$$

$$Rec = \frac{1}{c} \sum_{i=1}^{c} \frac{TP_i}{FN_i + TP_i}, \tag{27}$$

$$F1 = \frac{2 \times Pre \times Rec}{Pre + Rec}, \tag{28}$$

$$Spe = \frac{1}{c} \sum_{i=1}^{c} \frac{TN_i}{TN_i + FP_i}. \tag{29}$$

其中，准确率用于衡量模型整体分类正确的比例；精确率表示模型预测为正类样本中真实为正类的比例；召回率表示所有真实正类样本中

被模型正确识别的比例；F1 Score为精确率与召回率的调和平均，用于综合评价模型的分类能力；特异性用于衡量模型正确识别负类样本的能力。

此外，为了进一步评估模型的计算复杂度和推理效率，本文引入参数量（Parameters，Params）和浮点运算量（Floating Point Operations，FLOPs）作为模型复杂度指标。其中，Params（M）表示模型中所有可学习参数的总数量，单位为Million（M），用于衡量模型的存储开销；GFLOPs表示模型在单次前向推理过程中所需执行的浮点运算次数，单位为Giga FLOPs，用于衡量模型的计算复杂度。通常情况下，参数量和GFLOPs越小，说明模型在保持性能的同时具有更高的计算效率。模型参数量的计算方式为：

$$Params=\sum_{l=1}^{L}N_l, \tag{30}$$

其中：L表示网络层数，N_l表示第l层的可学习参数数量。

模型的浮点运算量表示为各层浮点运算次数的总和，其计算方式为：

$$FLOPs_{\text{total}}=\sum_{l=1}^{L}FLOPS_l, \tag{31}$$

其中：L表示网络层数，S_l表示第l层的浮点运算次数。

5.3 实验环境与参数设置

本实验基于Windows Server 2019 Datacenter 64位操作系统，硬件配置为Intel Xeon Gold 6154 CPU（36核，3.0 GHz）、256 GB内存及双路TITAN V GPU加速计算。模型开发采用Python语言，基于PyTorch（GPU版本）深度学习框架进行网络构建与训练。优化器选用AdamW，初始学习率设置为2×10^{-4}，权重衰减设为0.05；采用Warmup＋余弦退火复合策略调整学习率，其中前15个轮次为学习率热身阶段，随后按余弦曲线衰减，最终衰减至初始值的5%（即1×10^{-5}）。模型在肺部X射线数据集上训练200个轮次（epochs），批处理大小（batch size）设为64，损失函数采用带标签平滑的交叉熵损失（Label Smoothing Cross-Entropy Loss，平滑因子0.1）。具体实验设置如表1所示。

表1 实验设置

Tab. 1 Experimental setting

Training config	224×224
Optimizer	AdamW
Batch size	64
Epoch	200
Initial learning rate	2.00×10^{-4}
Learning rate factor	0.05
Learning rate schedule	LambdaLR＋Warmup＋Cosine
Loss	CrossEntropyLoss＋label_smoothing

5.4 对比实验

为了评估本文提出的QSMFormer模型在不同医学图像分类任务中的识别性能与适应性，本文在两个具有代表性的公开医学影像数据集上开展了对比实验，包括胸部X光图像数据集、用于眼科疾病检测和分类的彩色眼底图像数据集。对比实验选取了六个近年来具有代表性的Transformer或基于token选择的经典模型作为基线，包括ViT[2]，Swin-T[22]，DynamicViT[8]，EViT[10]，ToMe[9]，和GTPViT[23]，涵盖了从标准Vision Transformer到多尺寸交互、token选择与重组等不同方向的方法。实验在统一的设置下分别对比了七个指标（Accuracy，F1-score，Recall，Precision，Specificity，Param和Flops），其中加粗结果表示该指标下的最优性能。通过实验比较不同方法在两个公开数据集上的分类性能与计算复杂度。

5.4.1 在Kaggle胸部X光数据集上的对比实验的结果

为了验证QSMFormer模型的识别性能，将其与经典模型、近年来基于token选择的方法以及基于Transformer的模型在数据集1上进行对比，实验结果如表2所示（加粗字体表示各指标下的最优结果）。与对比模型相比，本文提出的QSMFormer在胸部X光数据集上取得了最佳的识别效果，五项性能指标（Accuracy，F1-score，Recall，Precision，Specificity）分别达到98.51%，97.02%，97.02%，97.03%和99.00%，表明该模型能够更充分地保留具有判别性的token，并减少冗余背景token对分类结果的干扰，从而提升识别性能。此外，QSMFormer的参数量为

表 2　基于胸部 X 光数据集的对比实验结果

Tab. 2　Comparative experimental results based on the chest X-ray dataset

Method	Acc /%	F1/%	Rec/%	Pre/%	Spe/%	Params/M	GFLOPs
ViT	95.52	91.04	91.03	91.05	97.02	85.80	16.85
SwinT	96.55	93.11	93.10	93.15	97.70	86.75	15.17
DynamicViT	97.57	95.14	95.15	95.16	98.38	86.76	13.63
EViT	95.90	91.79	91.79	91.81	97.26	85.80	13.65
ToMe	96.74	93.48	93.47	93.52	97.82	85.80	**11.70**
GTPViT	95.62	91.24	91.22	91.38	97.08	85.80	16.86
QSMFormer	**98.51**	**97.02**	**97.02**	**97.03**	**99.00**	**85.80**	12.35

85.80 M,计算量为 12.35 G,在保持较高分类性能的同时降低了推理计算量,体现出较好的性能与效率平衡。

5.4.2　Sharmin 9+1 数据集上的对比实验的结果

为了验证该结构在更广范围内的泛化能力,本文在另一个公开数据集上进行了对比实验。实验结果如表 3 所示(加粗字体表示各指标下的最优结果)。本文提出的 QSMFormer 在 Sharmin 9+1 数据集上取得了最佳的识别效果,五项性能指标(Accuracy,F1-score,Recall,Precision,Specificity)分别达到 98.26%,92.60%,93.59%,91.81% 和 98.99%。QSMFormer 在获得较高分类性能的同时保持了较低的计算复杂度。这表明本文方法在眼底图像分类任务中同样保持了较好的分类性能与计算效率。

表 3　基于 Sharmin 9+1 数据集的对比实验结果

Tab. 3　Comparative experimental results based on the ten-class fundus image dataset

Method	Acc /%	F1/%	Rec/%	Pre/%	Spe/%	Params/M	GFLOPs
ViT	97.86	90.63	91.61	89.78	98.76	85.81	16.85
SwinT	97.91	91.84	92.20	91.58	98.77	86.75	15.17
DynamicViT	98.08	92.11	92.58	91.74	98.88	86.77	13.63
EViT	97.86	91.10	91.66	90.66	98.75	85.81	13.65
ToMe	97.87	91.28	91.97	90.71	98.76	85.81	**11.70**
GTPViT	97.82	90.54	91.30	89.85	98.73	85.81	16.86
QSMFormer	**98.26**	**92.60**	**93.59**	**91.81**	**98.99**	**85.81**	12.35

5.4.3　在外部胸部 X 线测试集上的补充验证结果

为进一步验证所提方法在不同来源医学图像上的泛化能力,本文补充引入一组来自广州妇女儿童医学中心的胸部 X 线图像作为外部测试集。该测试集仅用于模型外部测试,不参与模型训练和参数调优。考虑到本文在数据集 1 上训练的模型标签空间为 COVID,Lung Opacity,Normal 和 Viral Pneumonia,而外部测试集中包含 COVID-19,Normal,Pneumonia 和 Tuberculosis 四类图像,本文选取其中可与训练标签空间对应或相近的 COVID-19,Normal 和 Pneumonia 三类进行测试。其中,COVID-19 和 Normal 分别映射为训练标签中的 COVID 和 Normal;Tuberculosis 类由于训练模型中无对应输出类别,未纳入本次测试。

需要说明的是,外部测试集中 Pneumonia 类未进一步区分病毒性肺炎和细菌性肺炎,而本文训练模型中的对应疾病类别为 Viral Pneumonia。考虑到二者均属于肺炎相关胸部 X 线异常表现,且本实验的目的在于评估模型在相近疾病类别和不同数据来源下的泛化能力,本文将外部测试集中的 Pneumonia 类近似映射为训练标签中的 Viral Pneumonia 类进行测试。为便于表述,后文将该映射类别简称为 Pneumonia。由于该映射并

非完全等价，本文仅将该实验作为补充性外部验证。测试过程中，所有模型均保持原4类输出不变；若样本被预测为Lung Opacity，则按误分类处理。最终共选取730张外部胸部X线图像用于测试，其中COVID-19，Normal和Pneumonia分别为106，234和390张。

表4　外部胸部X线测试集上的补充对比实验结果

Tab. 4　Supplementary comparative experimental results on the external chest X-ray test set

Method	Acc	F1	Rec	Pre	Spe	Params/M	GFLOPs
ViT	81.10	82.11	81.64	82.67	91.24	85.80	16.85
SwinT	83.15	82.00	79.89	86.18	90.33	86.75	15.17
DynamicViT	91.64	92.03	90.62	94.14	95.65	86.76	13.63
EViT	83.42	84.73	84.63	84.95	92.35	85.80	13.65
ToMe	89.18	90.01	88.68	91.51	95.04	85.80	**11.70**
GTPViT	81.37	83.00	82.55	83.82	91.82	85.80	16.86
QSMFormer	**93.84**	**94.40**	**92.90**	**96.10**	**97.21**	**85.80**	12.35

如表4所示，QSMFormer在外部胸部X线测试集上取得了最佳分类结果，其Accuracy，F1，Recall，Precision和Specificity分别达到93.84%，94.40%，92.90%，96.10%和97.21%。与次优方法DynamicViT相比，QSMFormer在Accuracy和F1上分别提升2.20%和2.37%；与标准ViT相比，Accuracy和F1分别提升12.74%和12.29%。该结果表明，在外部测试数据未参与训练和参数调优的情况下，QSMFormer仍能在不同来源胸部X线图像上取得相对稳定且优于对比方法的分类性能，说明所提出的区域分裂-合并式token选择策略具有较好的外部适应性和泛化潜力。由于外部测试集与训练数据在类别定义和数据来源上存在差异，该实验主要用于补充验证模型在外部数据上的稳定性，而非作为严格同分布测试。

5.4.4　边缘背景扰动下的鲁棒性实验

为进一步验证模型对非主体背景区域变化的稳定性，本文在胸部X光图像数据集的验证集上构建边缘背景扰动测试集。所有模型均使用原始训练集进行训练，不使用扰动图像参与训练或参数调优，仅在测试阶段分别输入原始验证集和扰动验证集，以比较不同方法在背景变化条件下的性能变化。

具体而言，本文对验证集中每幅图像四周边缘区域进行扰动处理。考虑到胸部X光图像的诊断主体通常位于图像中心区域，本文保持中心胸部主体区域不变，仅对图像四周宽度为原图尺寸8%的外圈区域采用该区域平均灰度值进行填充，以模拟黑边、边缘背景及非诊断区域变化对模型预测结果的影响。该实验能够从测试阶段扰动的角度评估模型是否过度依赖边缘背景信息。若模型在扰动后性能下降较小，则说明其对非主体背景变化具有更好的稳定性。

本文选取基础视觉Transformer（Vision Transformer，ViT）、层级视觉Transformer（Swin Transformer，SwinT）、典型动态token剪枝方法DynamicViT和典型token合并方法ToMe作为代表性对比模型，并与本文提出的QSMFormer进行比较。实验结果如表5所示。

由表5可见，在边缘背景扰动后，ViT和ToMe的性能下降较为明显。其中，ViT的Mean Acc和F1分别下降4.85%和9.54%，ToMe的Mean Acc和F1分别下降5.04%和10.29%，说明在本实验设置下，常规全局注意力模型和部分token合并方法对边缘背景变化更敏感。SwinT和DynamicViT在扰动条件下表现较稳定，但其扰动后的Mean Acc和F1仍低于QSMFormer。

相比之下，QSMFormer在扰动验证集上仍取得最高的Mean Acc和F1，分别为97.29%和94.59%。与原始验证集结果相比，QSMFormer的Mean Acc仅下降1.21%，F1下降2.43%，表

表 5　边缘背景扰动下不同方法的鲁棒性比较

Tab. 5　Robustness comparison of different methods under edge-background perturbation　　(%)

Method	Clean Mean Acc	Perturbed Mean Acc	Mean Acc Drop	Clean F1	Perturbed F1	F1 Drop
ViT	95.52	90.67	4.85	91.04	81.50	9.54
SwinT	96.55	95.43	1.12	93.12	90.89	2.23
DynamicViT	97.57	96.27	1.31	95.14	92.58	2.56
ToMe	96.74	91.70	5.04	93.48	83.19	10.29
QSMFormer	98.51	97.29	1.21	97.02	94.59	2.43

明所提出的四叉树 token 分裂-合并机制在保持较高分类性能的同时,对边缘背景变化具有较好的稳定性。该结果说明 QSMFormer 能够更加关注胸部 X 光图像中的主体结构信息,并在一定程度上降低模型对非主体背景区域的依赖。

5.5　消融实验

5.5.1　区域分裂合并模块嵌入位置的消融研究

为了研究区域分裂合并模块在 Transformer 网络中的最优嵌入位置,本文在基准模型的 12 层 Transformer 架构中进行位置消融实验。具体而言,将区域分裂合并模块分别插入第 i 个 Transformer 块之后(i=0,1,…,11),并保持其余训练设置完全一致,从而系统评估不同嵌入位置对模型性能的影响。实验评价指标包括 Accuracy, F1 Score, Recall, Precision, Specificity 以及模型复杂度指标 Params 和 GFLOPs。不同嵌入位置的实验结果如表 6 所示。

从实验结果可以观察到,模型性能随着剪枝位置的变化呈现出明显的“先升后降”趋势。当区域分裂合并模块位于浅层时(如第 0 层),模型性能相对较低(Acc 为 95.52%),这可能是由于网络尚未充分提取有效特征,过早进行 token 剪枝容易破坏基础特征表示。随着模块逐渐向中层移动,模型性能持续提升,并在第 6 层达到最佳结果(Acc 为 98.51%,F1 Score 为 97.02%),表明此时 token 选择操作能够在保留关键语义信息的同时减少冗余 token,从而实现性能与效率之间的良好平衡。当剪枝位置进一步向深层移动时,模型性能开始下降,例如在第 7 层 Acc 降至 97.85%。虽然在第 8~10 层性能有所回升,但整体仍未超过第 6 层的峰值,而在第 11 层再次出现下降。这一现象表明,在深层阶段进行剪枝可能导致重要语义信息的丢失,从而影响模型性能。

表 6　QSMFormer 在不同层数的性能差距

Tab. 6　Performance gap of QSMFormer with different numbers of layers

Prune_at_block	Acc/%	F1/%	Rec/%	Pre/%	Spe/%	Params/M	GFLOPs
0	95.52	91.04	91.04	91.29	97.01	85.8	6.96
1	96.83	93.66	93.66	93.76	97.88	85.8	7.86
2	97.39	94.78	94.78	94.78	98.26	85.8	8.76
3	97.76	95.52	95.53	95.54	98.51	85.8	9.66
4	98.13	96.27	96.28	96.28	98.76	85.8	10.56
5	98.32	96.65	96.64	96.68	98.88	85.8	11.45
6	**98.51**	**97.02**	**97.02**	**97.03**	**99.00**	**85.8**	12.35
7	97.85	95.71	95.72	95.72	98.57	85.8	13.25
8	98.23	96.47	96.46	96.47	98.82	85.8	14.15
9	98.23	96.46	96.47	96.48	98.82	85.8	15.05
10	98.32	96.65	96.65	96.67	98.88	85.8	15.95
11	97.67	95.35	95.34	95.39	98.44	85.8	16.85

从计算复杂度的角度来看，随着剪枝位置的后移，模型的GFLOPs近似呈线性增长（由6.96增加至16.85），而参数量始终保持不变（85.8 M），说明区域分裂合并模块本身不会引入额外的可学习参数，但不同嵌入位置会影响后续计算阶段的token数量，从而改变整体计算开销。综合考虑模型性能与计算效率，第6层在保持最高分类性能的同时具有适中的计算复杂度（12.35 GFLOPs），因此可以认为是当前模型结构下区域分裂合并模块的最优嵌入位置。这一结果与Transformer的特征学习规律相一致：浅层主要学习局部纹理和边缘等低级特征，中层逐渐形成更具语义性的结构表示，而深层则建模全局语义关系。因此，在中层阶段进行token选择能够在信息保留与冗余去除之间取得更好的平衡，从而提升模型整体性能。

5.5.2 关键参数τ与k_r对QSM方法性能影响的消融研究

区域分裂合并模块包含两个关键超参数：分裂阈值τ与区域保留比例kr。其中，τ控制区域分裂的粒度，而k_r决定每个区域中保留token的比例，从而共同影响模型的token压缩程度与特征表达能力。为研究两者对模型性能的影响，本文在最佳剪枝位置（第6层）下进行网格搜索实验，其中$\tau\in\{0.6,0.65,0.7,0.75\}$，$k_r\in\{0.25,0.3,0.35,0.4\}$，共16组参数组合。实验采用Accuracy和F1 Score作为主要评价指标，其余训练设置保持一致。实验结果如表7所示。

表7 不同τ与k_r参数组合下QSM方法性能（数值表示为Acc（%）/F1（%））

Tab. 7 Performance of the QSM method under different combinations of τ and k_r parameters (values are expressed as Acc (%)/F1(%))

τ/k_r	0.25	0.3	0.35	0.4
0.6	97.67/95.53	98.04/96.08	98.22/96.46	97.76/95.53
0.65	97.76/95.52	97.76/95.52	**98.51/97.02**	98.22/96.46
0.7	98.04/96.09	98.41/96.83	97.95/95.90	97.94/95.89
0.75	98.13/96.28	97.85/95.71	97.66/95.34	97.76/95.53

实验结果表明，QSM方法在不同参数组合下整体表现稳定，其中$\tau=0.65$与$k_r=0.35$的组合取得最佳性能（Acc为98.51%，F1为97.02%）。从整体趋势来看，模型性能在$\tau\in[0.65,0.7]$和$k_r\in[0.3,0.35]$区间内表现较为稳定。当τ取值过小时，区域容易被过度分裂，导致区域结构碎片化并削弱区域级语义一致性；而当τ过大时，区域分裂不足，难以有效区分重要区域与冗余区域。另一方面，k_r过小会导致区域内信息保留不足，而过大的k_r又会降低token压缩效果，从而影响模型性能。因此，两者需要协同调节以在信息保留与冗余削减之间取得平衡。

综合来看，QSM方法在合理参数范围内具有较好的稳定性，其中$\tau=0.65$与$k_r=0.35$在当前实验设置下取得最佳综合性能。该结果表明，适度的区域分裂粒度与合理的token保留比例能够有效提升区域级特征表达能力，从而进一步发挥区域分裂合并策略在token选择中的优势。

5.5.3 模块消融实验

为进一步验证QSMFormer中各模块的有效性，本文在相同实验设置下对模型进行了模块级消融实验。考虑到空间与语义耦合的异质性评估模块直接服务于后续四叉树递归分裂过程，二者共同决定区域的自适应生成方式，因此本文将其作为“异质性驱动分裂”整体进行消融。在w/o Heterogeneity-guided Split变体中，移除基于区域异质性阈值的自适应分裂机制，并采用固定4×4区域划分作为替代，以保证后续区域合并与区域内Top-K选择流程仍可正常执行。此外，本文分别构建w/o Merge和w/o Top-K两个变体，用于评估区域合并模块和区域内Top-K Token选择策略对模型性能的影响。各消融变体的模块组成如表8所示。

由表9可见，完整QSMFormer在各项分类指标上均取得最优结果，Acc，F1，Rec，Pre和Spe分别达到98.51%，97.02%，97.02%，97.03%和99.00%。与ViT baseline相比，QSMFormer的

表 8　QSMFormer 模块消融实验设置

Tab. 8　Ablation settings of modules in QSMFormer

Method	Heterogeneity-guided split	Merge	Top-K
ViT baseline	×	×	×
w/o Heterogeneity-guided Split	×	√	√
w/o Merge	√	×	√
w/o Top-K	√	√	×
QSMFormer	√	√	√

Acc 提升 2.99%，F1 提升 5.98%，同时 GFLOPs 由 16.85 降低至 12.35，说明所提出的 token 压缩机制在提升识别性能的同时降低了主干网络的计算开销。此外，各模型的参数量均为 85.80 M，说明所提出模块未引入额外可学习参数。

从各消融变体来看，当移除异质性驱动分裂机制后，模型 Acc 和 F1 分别下降至 97.57% 和 95.16%，表明基于区域异质性的自适应区域生成有助于模型根据图像内容构建更合理的区域结构。当移除区域合并模块后，Acc 和 F1 进一步下降至 97.48% 和 94.97%，说明仅进行区域分裂容易造成区域碎片化，而区域合并能够整合相邻且语义一致的区域，从而增强区域级表达的稳定性。当移除区域 Top-K 选择策略后，模型 Acc 和 F1 分别为 98.04% 和 96.09%，仍低于完整 QSMFormer，说明区域内关键 token 选择有助于进一步保留高判别性信息并减少冗余背景 Token 的干扰。综上，异质性驱动分裂、区域合并和区域 Top-K 选择三个模块均对 QSMFormer 的最终性能具有积极贡献。

表 9　QSMFormer 模块消融实验结果

Tab. 9　Ablation results of modules in QSMFormer

Method	Acc/%	F1/%	Rec/%	Pre/%	Spe/%	Params/M	GFLOPs
ViT baseline	95.52	91.04	91.03	91.05	97.02	85.80	16.85
w/o Heterogeneity-guided Split	97.57	95.16	95.17	95.21	98.38	85.80	12.35
w/o Merge	97.48	94.97	94.97	95.01	98.32	85.80	12.35
w/o Top-K	98.04	96.09	96.09	96.10	98.69	85.80	12.35
QSMFormer	98.51	97.02	97.02	97.03	99.00	85.80	12.35

6　结　论

在医学图像计算机辅助诊断中，模型容易受到背景噪声、非诊断相关区域以及数据采集差异的影响，从而降低分类性能与稳定性。针对医学图像分类任务中背景冗余信息干扰的问题，本文提出了一种基于区域分裂与合并机制的 Transformer 框架 QSMFormer。该方法通过在中间层引入区域感知的 token 选择策略，降低冗余背景信息对分类决策的干扰，在提升分类性能的同时减少推理计算量。

具体而言，QSMFormer 通过对 token 特征进行空间重组与结构化建模，实现从“逐 token 独立建模”向“区域感知建模”的转变。首先，模型基于区域内部特征异质性对 token 空间进行递归分裂，使语义变化显著或结构复杂的区域获得更细粒度的表达，从而充分保留关键的诊断线索；随后，结合区域间嵌入相似性与 token 显著性分数的一致性，对相邻区域进行合并，以缓解过度分裂带来的碎片化问题，并构建语义一致且结构稳定的区域表示。在此基础上，通过区域配额分配与区域内 top-K 选择策略，模型能够在保证空间多样性的同时，筛选出最具诊断价值的 token，从而在压缩计算开销的同时提升特征表达的判别性与结构稳定性。在一个公开胸部 X 射线数据集和 Sharmin 9+1 数据集上进行验证，其准确率、F1、召回率、精确率、特异度在胸部 X 射线数据集上分别达到 98.51%，97.02%，97.02%，97.03% 和 99.00%，在 Sharmin 9+1 数据集上分别达到 98.26%，92.60%，93.59%，91.81% 和 98.99%。

总体而言，本文从区域结构建模的角度为降

低医学图像分类模型的背景依赖提供了一种新的思路。然而，基于Transformer的去偏见研究仍面临诸多挑战，未来可从以下几个方向进一步探索：

(1)背景干扰与潜在伪相关问题具有显著的数据依赖性，其表现形式与强度在不同疾病类型、人群分布及采集环境中存在差异。未来可在更多跨中心、多模态医学数据集上验证QSMFormer的泛化能力与稳定性。

(2)本文主要从token选择与信息组织的角度缓解背景干扰及潜在伪相关影响，但尚未在严格因果推断框架下对虚假关联进行建模。未来可结合因果推理方法，通过构建结构化因果模型区分因果特征与伪相关特征，从而进一步提升模型的可解释性与可靠性。

(3)在模型设计方面，本文基于固定尺寸的patch表示进行token建模，并通过区域分裂与合并机制实现多尺度区域构建。未来可进一步探索从输入阶段引入多尺度或patch划分策略，以更直接地建模不同尺度的结构信息。

作者贡献声明：

周涛：方法的整体构思和设计；
沈飞宇：论文撰写和实验实施；
刘洋：测量实验数据分析；
常德芳：测量实验数据管理；
于诗怡：测量实验数据验证；
陈铮：实验材料提供。

参考文献：

[1] 赵阳，李俊诚，成博栋，等. 深度学习在口腔医学影像中的应用与挑战[J]. 中国图象图形学报，2024，29(3)：586-607.
ZHAO Y，LI J C，CHENG B D，*et al*. Applications and challenges of deep learning in dental imaging[J]. *Journal of Image and Graphics*，2024，29(3)：586-607.(in Chinese)

[2] DOSOVITSKIY A，BEYER L，KOLESNIKOV A，*et al*. An Image Is Worth 16x16 Words：Transformers for Image Recognition at Scale[EB/OL]. 2020：*arXiv*：2010.11929. https://arxiv.org/abs/2010.11929

[3] 周涛，彭彩月，杜玉虎，等. DRT Net：面向特征增强的双残差Res-Transformer肺炎识别模型[J]. 光学 精密工程，2024，32(5)：714-726.
ZHOU T，PENG C Y，DU Y H，*et al*. DRT Net：Dual res-transformer pneumonia recognition model oriented to feature enhancement[J]. *Opt. Precision Eng.*，2024，32(5)：714-726.(in Chinese)

[4] 邢素霞，鞠子涵，刘子骄，等. 视觉Transformer预训练模型的胸腔X线影像多标签分类[J]. 中国图象图形学报，2023，28(4)：441.
XING S X，JU Z H，LIU Z J，*et al*. Multi-label classification of chest X-ray images with pre-trained vision Transformer model[J]. *Journal of Image and Graphics*，2023，28(4)：441.(in Chinese)

[5] YOU C Y，DAI H C，MIN Y F，*et al*. Uncovering memorization effect in the presence of spurious correlations[J]. *Nature Communications*，2025，16：5424.

[6] ZHOU T，NIU Y X，LU H L，*et al*. Transformer optimization algorithm for selecting tokens based on genetic algorithm[J]. *Applied Soft Computing*，2026，190：114632.

[7] ZHOU T，NIU Y X，LU H L，*et al*. Vision transformer：To discover the “four secrets” of image patches[J]. *Information Fusion*，2024，105：102248.

[8] RAO Y M，ZHAO W L，LIU B L，*et al*. DynamicViT：Efficient Vision Transformers with Dynamic Token Sparsification[EB/OL]. 2021：*arXiv*：2106.02034. https://arxiv.org/abs/2106.02034

[9] BOLYA D，FU C Y，DAI X L，*et al*. Token Merging：Your ViT but Faster[EB/OL]. 2022：*arXiv*：2210.09461. https://arxiv.org/abs/2210.09461

[10] LIANG Y W，GE C J，TONG Z，*et al*. Not All Patches are What You Need：Expediting Vision Transformers *via* Token Reorganizations[EB/OL]. 2022：*arXiv*：2202.07800. https://arxiv.org/abs/2202.07800

[11] ZENG W，JIN S，XU L M，*et al*. TCFormer：visual recognition *via* token clustering transformer[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*，2024，46(12)：9521-9535.

[12] 李敏，王月，张玉川，等. 基于Vision Transformer的胸片多标签疾病分类模型[J]. 激光杂志，2024，45(9)：132-140.

LI M, WANG Y, ZHANG Y C, *et al*. Chest X-ray multi-label disease classification model based on vision transformer[J]. *Laser Journal*, 2024, 45(9): 132-140. (in Chinese)

[13] ZEHLIKE M, LOOSLEY A, JONSSON H, *et al*. Beyond incompatibility: Trade-offs between mutually exclusive fairness criteria in machine learning and law [J]. *Artificial Intelligence*, 2025, 340: 104280.

[14] KAMBLE T S, WANG H T, MYERS N, *et al*. Predicting cancer survival at different stages: Insights from fair and explainable machine learning approaches [J]. *International Journal of Medical Informatics*, 2025, 197: 105822.

[15] GONZÁLEZ-SENDINO R, SERRANO E, BAJO J. Quantifying algorithmic discrimination: a two-dimensional approach to fairness in artificial intelligence [J]. *Engineering Applications of Artificial Intelligence*, 2025, 144: 109979.

[16] GOCERI E. An efficient network with CNN and transformer blocks for glioma grading and brain tumor classification from MRIs [J]. *Expert Systems with Applications*, 2025, 268: 126290.

[17] SEYYED-KALANTARI L, ZHANG H R, MCDERMOTT M B A, *et al*. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations [J]. *Nature Medicine*, 2021, 27(12): 2176-2182.

[18] MARIN D, CHANG J R, RANJAN A, *et al*. Token pooling in vision transformers for image classification [C]. 2023 *IEEE/CVF Winter Conference on Applications of Computer Vision* (*WACV*). *January* 2-7, 2023, *Waikoloa*, *HI*, *USA*. IEEE, 2023: 12-21.

[19] HOROWITZ S L, PAVLIDIS T. Picture segmentation by a tree traversal algorithm[J]. *Journal of the ACM*, 1976, 23(2): 368-388.

[20] RAHMAN T, KHANDAKAR A, QIBLAWEY Y, *et al*. Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images[J]. *Computers in Biology and Medicine*, 2021, 132: 104319.

[21] SHARMIN S, RASHID M R, KHATUN T, *et al*. A dataset of color fundus images for the detection and classification of eye diseases [J]. *Data in Brief*, 2024, 57: 110979.

[22] LIU Z, LIN Y T, CAO Y, *et al*. Swin transformer: hierarchical vision transformer using shifted windows[C]. 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). *October* 10-17, 2021, *Montreal*, *QC*, *Canada*. IEEE, 2022: 9992-10002.

[23] XU X W, WANG S, CHEN Y D, *et al*. GTP-ViT: efficient vision transformers *via* graph-based token propagation [C]. 2024 *IEEE/CVF Winter Conference on Applications of Computer Vision* (*WACV*). *January* 3-8, 2024, *Waikoloa*, *HI*, *USA*. IEEE, 2024: 86-95.

作者简介:

周　涛(1977—),男,宁夏同心人,博士,教授,2010年于西北工业大学获得博士学位,主要从事医学图像分析处理、深度学习、模式识别等方面的研究。E-mail: zhoutaonxmu@126.com

通讯作者:

沈飞宇(1998—),男,宁夏银川人,硕士研究生,主要从事智能医学影像图像处理,计算机辅助诊断等方面的研究。E-mail: shenfeiyu16@stu.nmu.edu.cn